JOURNAL OF VIROLOGY, Apr. 1985, p. 65–71
0022-538X/85/040065-07$02.00/0

Vol. 54, No. 1

The Envelope-Associated 22K Protein of Human Respiratory Syncytial Virus: Nucleotide Sequence of the mRNA and a Related Polytranscript

PETER L. COLLINS†* AND GAIL W. WERTZ

Department of Microbiology and Immunology, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27514

Received 10 September 1984/Accepted 21 December 1984

We recently determined that respiratory syncytial virus (strain A2) encodes a fourth unique envelope-associated virion protein that has molecular weight of approximately 24,000, as estimated by gel electrophoresis. The nucleotide sequence of the mRNA encoding this novel protein has now been determined from five cDNA clones, including three that contain the complete mRNA sequence. The complete mRNA sequence is 957 nucleotides, exclusive of polyadenylate, and contains two partially overlapping open reading frames. The 5′-proximal open reading frame is favored for utilization by the criteria of the location and sequence of its translational start site. Furthermore, the calculated molecular weight of the encoded protein, 22,153, is in agreement with the previous estimate of 24,000 for the authentic protein identified by hybrid selection and in vitro translation. The sequence of the predicted protein, now designated the 22K protein, contains 194 amino acids, is relatively hydrophilic, and appears to be the most basic of the respiratory syncytial virus proteins. The mRNA also contains a second, internal open reading frame which would encode a protein of 90 amino acids. However, no evidence for this translation product is known. The first nine nucleotides in the mRNA sequence, 5′-GGGGCAAAU, are identical to the conserved sequence identified previously at the 5′ termini of seven other respiratory syncytial viral mRNAs; the sequence at the 3′ end of the 22K mRNA, 5′. . . AGUUAUUU-polyadenylate, contains the elements of the previously identified 3′-terminal consensus sequence for respiratory syncytial virus mRNAs, $AGU^{A}_{U}A(N)_{1-4}$-polyadenylate (P. L. Collins, Y. T. Huang, and G. W. Wertz, Proc. Natl. Acad. Sci. U.S.A. 81:7683–7687). In addition, we present and describe the intergenic sequence of a dicistronic RNA derived from readthrough of the F and 22K protein genes.

Human respiratory syncytial (RS) virus is an enveloped, cytoplasmic, RNA-containing paramyxovirus that is an important agent of respiratory tract disease (2). RS virus is classified in a separate subgroup, pneumovirus, because of differences in virion morphology and in the biological activities of the viral glycoproteins (21). More recently, as described below, determination of the RS virus genetic map and identification of RNA and protein gene products revealed differences in gene number and genome organization between the pneumovirus RS virus and other prototypic paramyxoviruses, such as Sendai virus (11, 12, 25) and Newcastle disease virus (3, 6).

The RS virus genome is a single negative strand of RNA and has a molecular weight of approximately 5.0×10^6 to 5.6×10^6 (4, 19). cDNA cloning experiments have identified 10 unique, nonoverlapping viral mRNAs (4, 5). Intracellular virus transcripts also include discrete species of polycistronic readthrough RNAs (4, 5), as has been demonstrated previously for the rhabdovirus vesicular stomatitis virus (15, 16) and for Newcastle disease virus (6, 40). Analyses of the RS virus polycistronic RNAs by Northern blotting (4, 5), together with UV mapping studies (7), determined the virus transcriptional map.

Translation in vitro of individual viral mRNAs isolated by gel electrophoresis (20) and by hybridization selection with cDNA clones (4) has shown that each mRNA encodes a single species of protein. The detection of additional minor proteins, possibly encoded by secondary reading frames (36, 39), has not been reproducible and remains to be investigated further. These experiments (4, 20) identified the RNA and protein products for the 10 viral genes. Seven of the 10 RS virus proteins have been identified unambiguously as virion structural proteins: these are the large protein (L; molecular weight, approximately 200,000); the major glycoprotein (G; molecular weight, 84,000); the fusion glycoprotein (F; molecular weight, 68,000); the major nucleocapsid protein (N; molecular weight, 42,000); the nucleocapsid phosphoprotein (P; molecular weight, 34,000); the matrix protein (M; molecular weight, 26,000); and the 22- to 25-kilodalton protein (22K) (10, 18, 30, 33, 34, 38, 41). Two viral proteins, the 11K and 14K proteins, appear to be nonstructural. The status as structural or nonstructural of the remaining known protein, the 9.5K protein, is uncertain (Y. T. Huang, P. L. Collins, and G. W. Wertz, Virus Res., in press). Six of the RS virus structural proteins, the L, G, F, N, P, and M proteins, have apparent counterparts among other paramyxoviruses. In contrast, the 22K protein lacks known counterparts. Recently, virion dissociation experiments in which nonionic detergent and increasing concentrations of salt were used identified the 22K protein as a component of the virus envelope, with solubility characteristics similar to those of the M protein (18).

As part of our efforts to characterize the structural components of the RS virus virion, we describe here the construction and sequencing of full-length cDNA clones of the mRNA encoding the 22K protein. The 22K protein had been previously designated VP25 (28, 30, 34, 41), VP6 (33), and

* Corresponding author.
† Present address: Laboratory of Infectious Diseases, Bldg. 7, Room 100, National Institute of Allergy and Infectious Diseases, Bethesda, MD 20205.

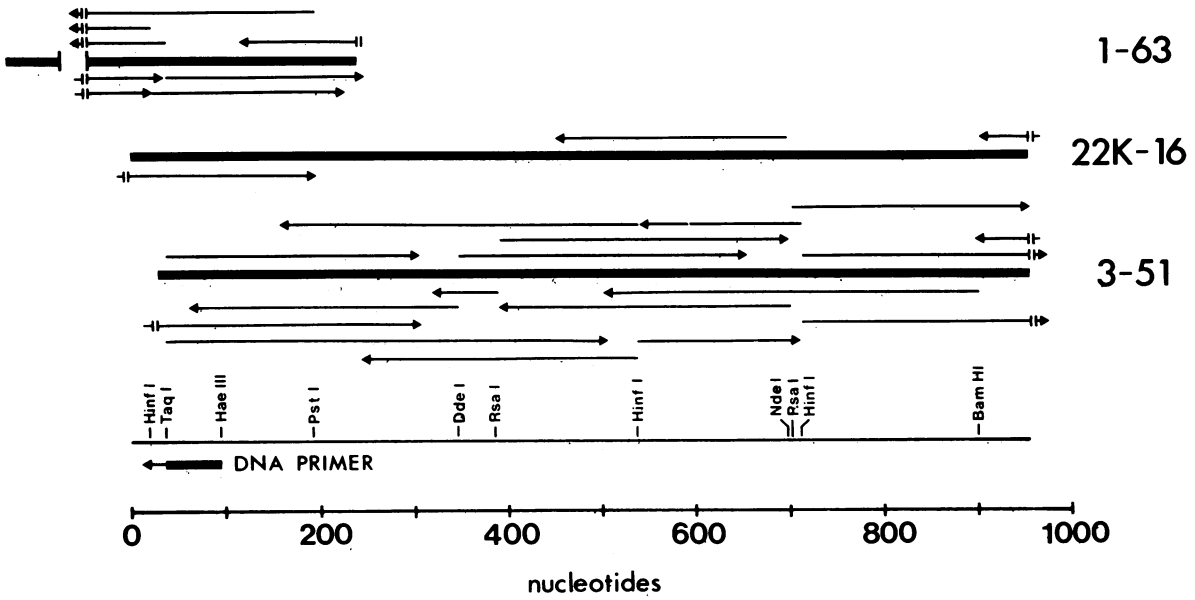

FIG. 1. Restriction site map and strategy for cDNA sequencing. cDNAs 1-63, 22K-16, and 3-51 are drawn to scale, except part of 1-63 representing sequences unrelated to the 22K mRNA has been omitted, as indicated by double slash marks. The orientation from left to right is 5′ to 3′ relative to the mRNA. The polydeoxyguanylate-polydeoxycytidylate tails and the polydeoxyadenylate tails at the 3′ ends of 22K-16 and 3-51 are omitted. The arrows indicate the directions and extents of sequence determination in the message-sense (arrows above lines) and antimessage strand (below lines). Double slash marks in certain arrows indicate where sequencing spanned the cDNA boundaries. Although not shown, two other full-length cDNAs, cDNAs 22K-14 and 22K-15, were sequenced to an extent similar to that of 22K-16. The map location of the DNA primer used for the primer extension experiments shown in Fig. 2 is indicated.

24K protein (4, 18). Here we use the designation 22K protein in accordance with the calculated molecular weight deduced from the nucleotide sequence.

MATERIALS AND METHODS

cDNA clones. All cDNAs were synthesized with intracellular polyadenylated RNA as template and were cloned in the *Pst*I site of the *Escherichia coli* plasmid pBR322 by polydeoxycytidylate · polydeoxyguanylate tailing. cDNAs 1-63 and 3-51 were constructed in earlier work (5). cDNAs 22K-14, 22K-15, and 22K-16 were isolated from a second cDNA library constructed under the following conditions, which were designed to maximize cDNA length (29). Polyadenylate [poly(A)]-selected RNA purified from eight 15-cm dishes of RS virus-infected HEp-2 cells was reverse transcribed in the presence of 0.8 mCi of [^{3}H]dCTP (specific activity, 400 mCi/mmol; ICN Radiochemicals) in an 800-μl reaction mixture under previously described conditions (5) with the addition of 80 μg of actinomycin D per ml. After incubation for 1 h at 43°C, the products were purified with phenol-chloroform and precipitated with ethanol. The nucleic acids were resuspended in water and incubated for 2 h at 37°C in the presence of 0.3 M NaOH (final volume, 300 μl). The mixture was neutralized by the additions of 25 μl of 2.5 M Tris-hydrochloride (pH 7.6) and 30 μl of 2 M HCl and was immediately passed through Sephadex G-200 with a column buffer of 1 mM Tris-hydrochloride (pH 7.6). The cDNAs contained in the leading edge of the void volume were collected, yielding 8.5 μg. Following the procedure of Land et al. (29), we added homopolymeric dCMP tails in a 550-μl reaction mixture containing 325 U of terminal transferase (P-L Biochemicals). The reaction mixture was incubated at 15°C. Aliquots were withdrawn after 2.5 and 5 min and adjusted to 10 mM EDTA, and the cDNAs were purified by extraction with phenol-chloroform, followed by three rounds of ethanol precipitation. Synthesis of the second cDNA strand was performed in a 600-μl reaction mixture under the conditions described above for reverse transcription of mRNA, except the actinomycin D was omitted, the oligodeoxythymidylate was replaced by 30 μg of oligodeoxyguanylate$_{12-18}$ (P-L Biochemicals) per ml, and the reaction contained 0.75 mCi of [α-^{32}P]dCTP (specific activity, 3,000 Ci/mmol; Amersham Corp.). After incubation for 1 h at 43°C, the reaction mixture was passed directly through Sepharose 6B, and the cDNAs in the void volume were recovered. To obtain maximum completion of second-strand synthesis, we assembled the cDNAs into a 400-μl reaction mixture containing 10 mM Tris-hydrochloride (pH 7.6), 8 mM magnesium acetate, 70 mM KCl, 10 mM dithioerythritol, 0.5 mM each deoxynucleotide, and 12 U of DNA polymerase I (Klenow fragment) (P-L Biochemicals). After incubation for 2 h at 15°C, the reaction was terminated by the addition of EDTA to 10 mM. The products were purified by extraction with phenol-chloroform and passed through Sepharose 6B, yielding 8 μg of double-stranded cDNA. Homopolymer dCMP tails were added in a 600-μl reaction mixture under the conditions described above, except incubation took place at 30°C for 2.5 and 5 min. The reactions were terminated by the addition of EDTA and by extraction with phenol-chloroform, and the products were collected by ethanol precipitation. The cDNAs were then electrophoresed on a 1.5% agarose gel under nondenaturing conditions in parallel with end-labeled size markers prepared from *Hpa*II digests of miscellaneous viral cDNA clones from the cDNA library previously described (5). The cDNAs and size markers were visualized by autoradiography, gel bands containing the appropriate cDNA size classes were excised, and the cDNAs were recovered by electroelution. Previously described conditions (5) were used for preparing *Pst*I-digested, oligodeoxyguanylate-tailed pBR322, for annealing plasmid and in-

serts, and for transforming competent cells of *Escherichia coli* HB101. cDNA clones of the 22K mRNA were identified by colony hybridization with nick-translated cDNA 3-51 (5) which had been purified from the vector by *Pst*I digestion and gel electrophoresis. These procedures followed conventional protocols (31).

DNA sequence analysis. End-labeled DNA fragments were prepared and analyzed by the chemical sequencing methods of Maxam and Gilbert (32) with several modifications. Before the chemical reactions took place, DNAs were purified by binding to and elution from NACS chromatography matrix (Bethesda Research Laboratories). For the adenine-plus-guanine reactions, DNA in 10 μl of water was mixed with 25 μl of formic acid (88%, reagent grade), incubated for 3 min at 20°C, and processed exactly as described for the hydrazine reactions (32). The cleavage products were analyzed on 5, 6, 8, and 20% polyacrylamide sequencing gels (0.35 mm by 25 by 40 to 100 cm). Except in the case of 20% gels, gels were fixed with 10% acetic acid and dried before autoradiography.

Primer extension. Recombinant plasmid 22K-16 (100 μg) was digested with *Hpa*II and *Hae*III, and the 179-base-pair fragment spanning the cDNA terminus from base-pair 99 in the 22K cDNA sequence to base pair 3660 in pBR322 was isolated, 5′ end labeled, digested with *Taq*I, and electrophoresed on a nondenaturing 10% polyacrylamide gel. The 57-base-pair, double-stranded *Taq*I-*Hae*III fragment, labeled at the 5′ end of the antimessage-sense strand, was recovered. For each experiment, approximately 20 pmol of primer (2×10^6 dpm) was denatured by boiling for 2 min, mixed with 20 μg of intracellular viral mRNA in a 50-μl mixture containing 80% formamide, 40 mM PIPES [piperazine-*N*,*N*′-bis(2-ethanesulfonic acid)] (pH 6.2), and 0.4 M NaCl, and then incubated for 10 min at 65°C, 10 min at 54°C, and 12 h at 42°C. The nucleic acids were precipitated twice with ethanol, and the hybridized primers were extended by reverse transcriptase in the presence or absence of dideoxynucleotides exactly as described previously (4a). After incubation, nucleic acids were precipitated with ethanol and analyzed on sequencing gels.

RESULTS AND DISCUSSION

cDNA sequencing. The restriction site map and strategy for sequencing cDNAs of the 22K mRNA are shown in Fig. 1. Most of the sequence was determined from the partial cDNAs 1-63 and 3-51, which were constructed in previous work (5). Sequencing was completed on three approximately full-length cDNAs, 22K-14, 22K-15, and 22K-16, synthesized as described above. The cDNA end that is 5′ relative to the mRNA was identified by primer extension experiments described below. The 3′ end was identified by the detection of terminal polydeoxyadenylate (13 to 23 residues) in cDNAs 3-51, 22K-14, and 22K-16. cDNA 22K-16 was found to contain the complete mRNA sequence. Interestingly, (see below), cDNA 1-63 was found to be a partial copy of a dicistronic readthrough RNA with the structure 5′ F mRNA–intergenic sequence–22K mRNA 3′.

Mapping the 5′ mRNA end. To precisely map the 5′ end of the 22K mRNA, we hybridized a 5′ end-labeled DNA primer (see above) to intracellular mRNA and extended to the 5′ mRNA end under conditions of partial chain termination (Fig. 2A). In previously published work (14), reverse transcriptase was shown to add an additional nucleotide to the ends of fully elongated primers, possibly owing to inefficient copying of the mRNA cap. A similar artifactual product (Fig. 2A, open arrow) was evident in the experiment shown in Fig. 2A and was discounted. The deduced sequence of the 5′ mRNA end was determined to be, in mRNA sense, 5′ NGGGCAAAU. The terminal nucleotide could not be identified by this method because all four dideoxynucleotide reactions terminate nonspecifically at the mRNA end. To identify the terminal nucleotide, we purified the fully elongated primer by gel electrophoresis and sequenced it by the chemical method (Fig. 2B). This established that the com-

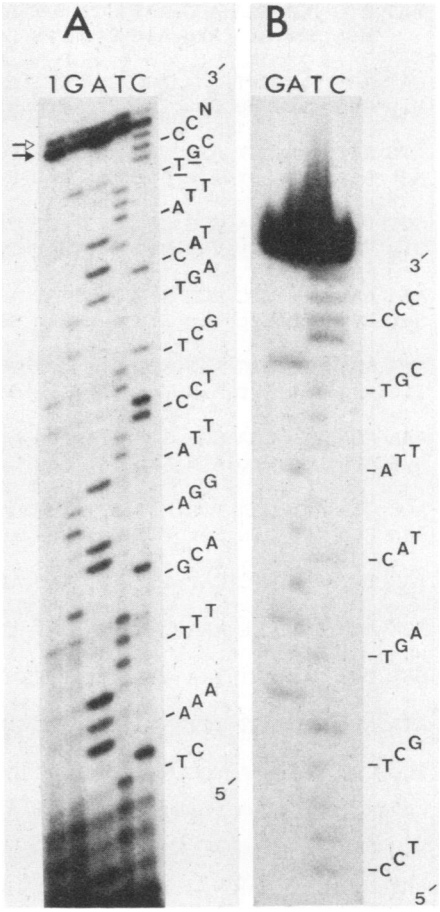

FIG. 2. Mapping and sequencing of the 5′ end of the 22K mRNA. (A) A 5′-end-labeled DNA primer was hybridized to intracellular mRNA and extended to the 5′ mRNA end by reverse transcriptase (lane 1). The position of the major extension product representing the authentic mRNA end is marked with a closed arrow; the open arrow indicates an artifactual end addition product described in the text. Lanes G, A, T, and C represent parallel reactions performed under conditions of partial chain termination with the appropriate dideoxynucleotide. The products were analyzed on an 8% sequencing gel. The deduced mRNA sequence in antimessage sense is shown. The fifth and sixth nucleotides (underlined) from the mRNA end could not be identified unambiguously by this experiment, but were identified as described in the legend to Fig. 2B. (B) The 5′ end-labeled primer was elongated on mRNA, and the extended product was purified by gel electrophoresis and sequenced by the chemical method. The cleavage products were analyzed on an 8% sequencing gel and visualized with an intensifying screen. The lane headings denote the reaction specificities: lane G, guanine; lane A, guanine plus adenine; lane T, cytosine plus thymine; lane C, cytosine. The deduced nucleotide sequence (antimessage sense) is shown.

```
GGG GCA AAT ATG TCA CGA AGG AAT CCT TGC AAA TTT GAA ATT CGA GGT CAT TGC TTA AAT GGT AAG AGG TGT CAT  75
          Met Ser Arg Arg Asn Pro Cys Lys Phe Glu Ile Arg Gly His Cys Leu Asn Gly Lys Arg Cys His  22

TTT AGT CAT AAT TAT TTT GAA TGG CCA CCC CAT GCA CTG CTT GTA AGA CAA AAC TTT ATG TTA AAC AGA ATA CTT 150
Phe Ser His Asn Tyr Phe Glu Trp Pro Pro His Ala Leu Leu Val Arg Gln Asn Phe Met Leu Asn Arg Ile Leu  47

AAG TCT ATG GAT AAA AGT ATA GAT ACC TTA TCA GAA ATA AGT GGA GCT GCA GAG TTG GAC AGA ACA GAA GAG TAT 225
Lys Ser Met Asp Lys Ser Ile Asp Thr Leu Ser Glu Ile Ser Gly Ala Ala Glu Leu Asp Arg Thr Glu Glu Tyr  72

GCT CTT GGT GTA GTT GGA GTG CTA GAG AGT TAT ATA GGA TCA ATA AAC AAT ATA ACT AAA CAA TCA GCA TGT GTT 300
Ala Leu Gly Val Val Gly Val Leu Glu Ser Tyr Ile Gly Ser Ile Asn Asn Ile Thr Lys Gln Ser Ala Cys Val  97

GCC ATG AGC AAA CTC CTC ACT GAA CTC AAT AGT GAT GAT ATC AAA AAG CTG AGG GAC AAT GAA GAG CTA AAT TCA 375
Ala Met Ser Lys Leu Leu Thr Glu Leu Asn Ser Asp Asp Ile Lys Lys Leu Arg Asp Asn Glu Glu Leu Asn Ser 122

CCC AAG ATA AGA GTG TAC AAT ACT GTC ATA TCA TAT ATT GAA AGC AAC AGG AAA AAC AAT AAA CAA ACT ATC CAT 450
Pro Lys Ile Arg Val Tyr Asn Thr Val Ile Ser Tyr Ile Glu Ser Asn Arg Lys Asn Asn Lys Gln Thr Ile His 147

CTG TTA AAA AGA TTG CCA GCA GAC GTA TTG AAG AAA ACC ATC AAA AAC ACA TTG GAT ATC CAT AAG AGC ATA ACC 525
Leu Leu Lys Arg Leu Pro Ala Asp Val Leu Lys Lys Thr Ile Lys Asn Thr Leu Asp Ile His Lys Ser Ile Thr 172

ATC AAC AAC CCA AAA GAA TCA ACT GTT AGT GAT ACA AAT GAC CAT GCC AAA AAT AAT GAT ACT ACC TGA CAA ATA 600
Ile Asn Asn Pro Lys Glu Ser Thr Val Ser Asp Thr Asn Asp His Ala Lys Asn Asn Asp Thr Thr ***       194

TCC TTG TAG TAT AAC TTC CAT ACT AAT AAC AAG TAG ATG TAG AGT TAC TAT GTA TAA TCA AAA GAA CAC ACT ATA 675

TTT CAA TCA AAA CAA CCC AAA TAA CCA TAT GTA CTC ACC GAA TCA AAC ATT CAA TGA AAT CCA TTG GAC CTC TCA 750

AGA ATT GAT TGA CAC AAT TCA AAA TTT TCT ACA ACA TCT AGG TAT TAT TGA GGA TAT ATA TAC AAT ATA TAT ATT 825

AGT GTC ATA ACA CTC AAT TCT AAC ACT CAC CAC ATC GTT ACA TTA TTA ATT CAA ACA ATT CAA GTT GTG GGA CAA 900

AAT GGA TCC CAT TAT TAA TGG AAA TTC TGC TAA TGT TTA TCT AAC CGA TAG TTA TTT                         957
```

FIG. 3. Complete nucleotide sequence of the 22K mRNA and the predicted protein sequence encoded by the 5′-proximal open reading frame.

plete sequence of the 5′ mRNA end is, in mRNA sense, 5′ GGGGCAAAU.

Nucleotide sequence of the 22K mRNA. The complete sequence of the 22K mRNA, exclusive of poly(A), contains 957 nucleotides (Fig. 3). The first nine nucleotides of the sequence, 5′ GGGGCAAAU, are identical to the consensus sequence that has been identified (4a) at the 5′ termini of the seven other RS virus mRNAs whose sequences are available: the G, 1A, 1B, and 1C mRNAs (P. L. Collins and G. W. Wertz, submitted for publication; unpublished data), F mRNA (4a), M mRNA (36), and N mRNA (9; unpublished results). In addition, the 3′ end of the 22K mRNA, 5′. . .AGUUAUUU-poly(A), is consistent with the consensus sequence, 5′. . .AGU$^{A}_{U}$A$(N)_{1-4}$ poly (A), previously identified for the seven mRNAs listed above (4a). Thus, as is the case for Sendai virus and vesicular stomatitis virus (11, 13, 14, 35, 37), the termini of RS virus genes contain conserved consequences. This contradicts the previous conclusion that RS virus mRNAs lack conserved termini (9, 36).

The 22K mRNA contains two major, partially overlapping open reading frames (Fig. 4). For eucaryotic mRNAs in general, a published survey of known translational start sites shows that (i) translation usually begins at the first, 5′-proximal AUG, and (ii) the presence of a purine, but not a pyrimidine, in the −3 position correlated positively with utilization (22–24). By both criteria, the translation start site for the 5′-proximal open reading frame, 5′-AAAUAUGU (nucleotides 6 through 13), has a structure that is consistent with utilization. This open reading frame encodes a polypeptide of 194 amino acids with a calculated molecular weight of 22,154 (Table 1). This is in good agreement with the molecular weight of 24,000 estimated by gel electrophoresis of the product identified by hybrid selection and in vitro translation (4). Thus, the available evidence indicates that

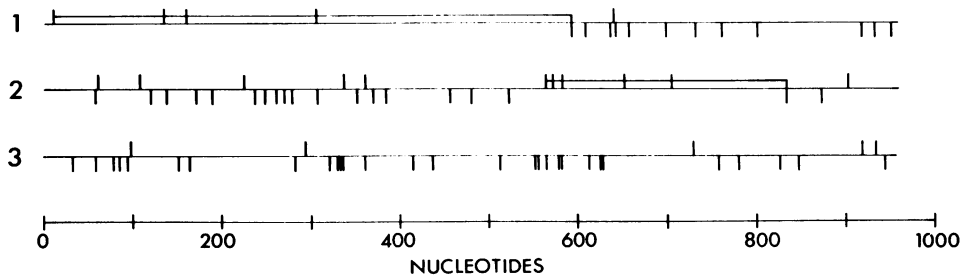

FIG. 4. Map of methionine codons (upper vertical lines) and translational stop codons (lower vertical lines) in the three reading frames of the 22K mRNA. The reading frames 1, 2, and 3 start with, respectively, the first, second, and third nucleotides in the sequence. The two longest open translational reading frames are marked with open rectangles.

TABLE 1. Amino acid compositon of the 194-amino acid (22, 154-dalton) protein product encoded by 22K mRNA

Amino acid	No. of residues
Alanine	8
Arginine	11
Asparagine	20
Aspartic acid	11
Cysteine	4
Glutamine	3
Glutamic acid	12
Glycine	6
Histidine	7
Isoleucine	16
Leucine	19
Lysine	18
Methionine	4
Phenylalanine	4
Proline	6
Serine	17
Threonine	13
Tryptophan	1
Tyrosine	5
Valine	9

the 5′-proximal open reading frame encodes the previously identified translation product of the 22K mRNA.

In contrast, the location and structure of the translational start site for the internal open reading frame, 5′. . .ACAAAUGA (nucleotides 559 through 566), are inconsistent with utilization. This open reading frame would encode a polypeptide of 90 amino acids, but there is presently no evidence that the internal open reading frame is utilized. If the internal open reading frame is not used, then 363 nucleotides at the 3′ end of the 22K mRNA are nontranslated. This would be unusual in comparison with the shorter nontranslated regions characteristic of other paramyxoviral, rhabdoviral, and influenza viral mRNAs sequenced to date. The possibility of an RNA splicing mechanism for expression of the internal open reading frame, as has been demonstrated for the M and NS mRNAs of influenza virus (26, 27), has not been ruled out but currently lacks experimental support because (i) RS virus transcription and replication occurs in the cytoplasm and appears to be independent of the cell nucleus, and (ii) truncated versions of the 22K mRNA were not detected in Northern blot experiments (5). Recently, it was shown that eucaryotic ribosomes that encountered a translational stop codon could reinitiate at a nearby, downstream translational start site (24). The demonstration of this reinitiation event suggested an alternate mechanism for expression of the internal open reading frame: ribosomes that terminate translation at nucleotides 592 through 594 at the end of the 5′-proximal open reading frame could reinitiate at nucleotides 650 through 652 and direct translation of the downstream two-thirds of the internal open reading frame to generate a polypeptide of 61 amino acids. No such protein has been reported, but a careful investigation has not been performed, and the reinitiation event might be infrequent but nonetheless authentic. These issues will require further investigation.

The two overlapping open reading frames described here for the 22K mRNA could have been generated artifactually if base deletions or insertions had occurred in the middle of a larger open reading frame during the synthesis or cloning of cDNA 3-51 or during the synthesis of the individual mRNA transcript represented by the clone. To investigate this possibility, we examined the sequence representing the overlap and flanking regions in two independent cDNAs, cDNAs 22K-15 and 22K-16 (Fig. 1). The sequencing spanned the overlap as well as several upstream translational stop codons in frames 2 and 3 and thus covered the region where the postulated insertion or deletion could have occurred. The results confirmed the sequence of cDNA 3-51. This provided evidence that the mRNA structure shown in Fig. 4 is authentic.

The 22K mRNA and its encoded protein have been identified only recently as unique viral gene products (4, 5). The deduced nucleotide and amino acid sequences presented here lack extensive homology with other available RS virus sequences, confirming that the 22K gene is unique and nonoverlapping.

Structure of the 22K protein. The amino acid sequence shown in Fig. 3 shows that the 194-amino acid 22K protein is rich in charged amino acids (59 residues), with a theoretical charge of +9.5 at neutral pH, assuming a contribution of +1 for arginine and lysine, +0.5 for histidine, and −1 for aspartate and glutamate. The prediction that the 22K protein is highly basic is in agreement with results from gel electrophoresis in the presence of a pH gradient (8). The predicted protein contains two clusters of basic amino acids: amino acids 3 through 48 include 13 basic and two acidic residues, and amino acids 139 through 162 include nine basic and one acidic residues. There are also two clusters of acidic residues: amino acids 51 through 71 include two basic and seven acidic residues, and amino acids 105 through 119 include three basic and six acidic residues. The remaining charged amino acids are scattered throughout the sequence.

The 22K protein has been shown to be associated with the virus envelope and to resemble the M protein in its sensitivity to solubilization with nonionic detergent and high salt (18). The exact nature, however, of the association of the 22K protein relative to the virion envelope remains to be established. To search for possible hydrophobic domains in the polypeptide chain, we prepared a plot of local hydrophobicity (17) versus sequence position (Fig. 5). This demonstrated that the 22K protein is relatively hydrophilic and contains only three moderately hydrophobic regions (amino acids 21 through 37, 72 through 91, and 127 through 135). Amino acids 72 through 91 are the best candidates for direct interaction with membrane: of 18 amino acids, 11 are hydro-

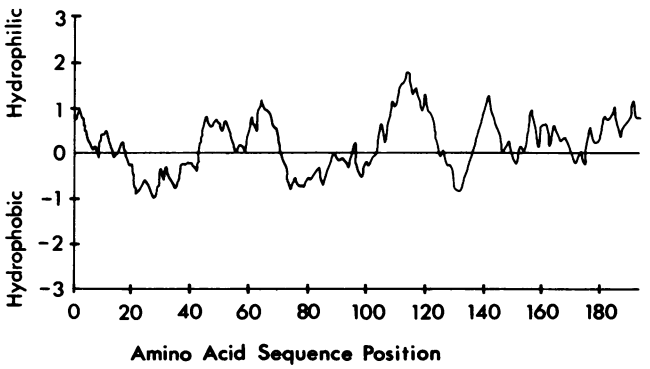

FIG. 5. Plot of hydrophilic (areas above line) and hydrophobic (areas below line) regions in the 22K protein. A window of 10 amino acids was used to calculate the local hydrophobicity of each position on the basis of published hydrophobicity values (17) (local hydrophobicity equals the average of the values for each group of 10 amino acids).

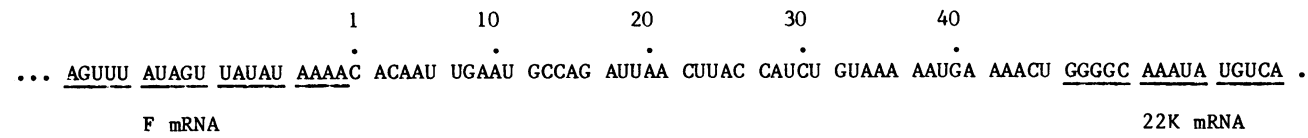

FIG. 6. Nucleotide sequences of the intergenic and flanking sequences in a polycistronic readthrough RNA. The sequence, written 5′ to 3′ (left to right) in mRNA sense, was deduced as described in the text by DNA sequencing of cDNA 1-63. Sequences that represent the 3′ end of the F mRNA (4a) and the 5′ end of the 22K mRNA (Fig. 3) are underlined. The intergenic sequence is numbered.

phobic and only 1 is charged. The 22K protein lacks an NH_2-terminal hydrophobic signal peptide and, in general, the amino acid sequence provides no compelling evidence for insertion into membranes.

The 22K protein was not labeled detectably in infected cells with mannose, glucosamine, or P_i (1, 28, 33), and the 22K protein extracted from infected cells or virions has the same electrophoretic mobility and peptide map as the 22K protein synthesized in vitro (4, 18). Thus, there is no evidence of post-translational processing.

Nucleotide sequence of a dicistronic RNA. cDNA 1-63 was previously identified as a cDNA of the F mRNA on the basis of high-stringency hybrid selection and translation in vitro (4). Subsequently, hybridization of cDNA 1-63 with Northern blots of intracellular viral RNA under conditions of low stringency showed that the clone hybridized strongly with the F mRNA, but also hybridized detectably with the 22K mRNA (data not shown). This suggested that cDNA 1-63 is a partial cDNA of a dicistronic transcript, which was confirmed by DNA sequencing and comparison with the nucleotide sequences of the F mRNA (4a) and 22K mRNA (Fig. 3). The sequence analysis showed that the message-sense strand of cDNA 1-63 contains 1,590 nucleotides representing the 3′ end of the F mRNA followed by four A residues, an intergenic sequence of 46 nucleotides, and 235 nucleotides representing the 5′ end of the 22K mRNA. The intergenic and flanking sequences are shown in Fig. 6. The 3′ end of the F mRNA and 5′ end of the 22K mRNA are represented in cDNA 1-63 exactly, without deletion or rearrangement. This supports the conclusion that the RNA represented by cDNA 1-63 is a readthrough transcript at the F and 22K genes and contradicts previous suggestions (9, 36) that the readthrough RNAs are transcribed from aberrant deletion mutants lacking regulatory signals at the gene boundaries. The structure of the dicistronic RNA, 5′ mRNA-intergenic sequence-22K mRNA, is in agreement with the viral transcriptional map (4, 5, 7).

We are presently sequencing intergenic regions in RS virus genomic RNA. However, the poor growth of RS virus in tissue culture and the instability of the virion have presented formidable obstacles to this analysis. Therefore, we have analyzed RNA polytranscripts as an alternate, concurrent approach to investigating intergenic sequences. Analogous polytranscripts had been analyzed previously for vesicular stomatitis virus (15, 16) and Sendai virus (4). For these two viruses, RNA sequencing showed that the intergenic sequences in the polytranscripts were exact complementary copies of the intergenic sequences in the genomic RNA, except that in the case of the vesicular stomatitis virus polytranscripts the intervening poly(A) tract was much longer than the seven residues encoded directly in the genome (14, 16). In the RS virus polytranscript analyzed here (Fig. 6), the upstream cistron is followed by four A residues. For four other polytranscripts analyzed to date, the upstream cistrons are followed by four to six A residues (data not shown). Thus, the RS virus polytranscripts lack the long intervening poly(A) tract characteristic of vesicular stomatitis virus polytranscripts (15, 16) and instead have short oligoadenylate tracts similar to those of Sendai virus (14), although unlike Sendai virus, the RS virus oligoadenylate tracks vary in length. The intergenic structure that follows the four A residues (Fig. 6) is surprisingly long, containing 46 nucleotides, and therefore has no apparent similarity to the dinucleotide and trinucleotide intergenic regions of vesicular stomatitis virus and Sendai virus, respectively. Surprisingly, the intergenic sequence shown in Fig. 6 also has no apparent resemblance to the four other sequenced RS virus polytranscript intergenic sequences, which contain 1, 9, 19, and 26 nucleotides (data not shown). Two of these intergenic sequences were determined by direct sequencing of intracellular RNA (data not shown), indicating that, despite the lack of sequence conservation among different polytranscripts, each polytranscript population is homogeneous. It will be important to extend these preliminary observations by direct analysis of genomic RNA.

ACKNOWLEDGMENTS

We thank Steve Langer for excellent technical assistance and John Glass and Andrew Ball for assistance with computer work.

This work was supported by Public Health Service grants AI12464 and AI20181 from the National Institute of Allergy and Infectious Diseases and training grant CA09156 from the National Cancer Institute. P.L.C. received additional support by grant AI06956 from the National Institute of Allergy and Infectious Diseases.

LITERATURE CITED

1. **Cash, P., C. R. Pringle, and C. M. Preston.** 1979. The polypeptides of human respiratory syncytial virus: products of cell-free protein synthesis and post-translational modifications. Virology **92**:375–384.
2. **Chanock, R. M., H. W. Kim, C. D. Brandt, and R. H. Parrott.** 1982. Respiratory syncytial virus, p. 471–489. *In* A. S. Evans (ed.), Viral infections in humans, epidemiology and control. Plenum Publishing Corp., New York.
3. **Collins, P. L., L. E. Hightower, and L. A. Ball.** 1980. Transcriptional map for Newcastle disease virus. J. Virol. **35**:682–693.
4. **Collins, P. L., Y. T. Huang, and G. W. Wertz.** 1984. Identification of a tenth mRNA of respiratory syncytial virus and assignment of polypeptides to the 10 viral genes. J. Virol. **49**:572–578.

4a. **Collins, P. L., Y. T. Huang, and G. W. Wertz.** 1984. Nucleotide sequence of the gene encoding the fusion (F) glycoprotein of human respiratory syncytial virus. Proc. Natl. Acad. Sci. U.S.A. **81**:7683–7687.

5. **Collins, P. L., and G. W. Wertz.** 1983. cDNA cloning and transcriptional mapping of nine polyadenylated RNAs encoded by the genome of human respiratory syncytial virus. Proc. Natl. Acad. Sci. U.S.A. **80**:3208–3212.
6. **Collins, P. L., G. W. Wertz, L. A. Ball, and L. E. Hightower.** 1982. Coding assignments of the five smaller mRNAs of Newcastle disease virus. J. Virol. **43**:1024–1031.
7. **Dickens, L. E., P. L. Collins, and G. W. Wertz.** 1984. Transcriptional mapping of human respiratory syncytial virus. J. Virol. **52**:364–369.
8. **Dubovi, E. J.** 1982. Analysis of proteins synthesized in respiratory syncytial virus-infected cells. J. Virol. **42**:372–378.

9. **Elango, N., and S. Venkatesan.** 1983. Amino acid sequence of respiratory syncytial virus capsid protein. Nucleic Acids Res. **11**:5941–5951.
10. **Fernie, B. F., and J. L. Gerin.** 1982. Immunochemical identification of viral and nonviral proteins of the respiratory syncytial virus virion. Infect. Immun. **37**:243–249.
11. **Giorgi, C., B. M. Blumberg, and D. Kolakofsky.** 1983. Sendai virus contains overlapping genes expressed from a single mRNA. Cell **35**:829–836.
12. **Glazier, K., R. Raghow, and D. W. Kingsbury.** 1977. Regulation of Sendai virus transcription: evidence for a single promoter in viro. J. Virol. **21**:863–871.
13. **Gupta, K. C., and D. W. Kingsbury.** 1982. Conserved polyadenylation signals in two negative-strand RNA virus families. Virology **120**:518–523.
14. **Gupta, K. C., and D. W. Kingsbury.** 1984. Complete sequences of the intergenic and mRNA start signals in the Sendai virus genome—homologies with the genome of vesicular stomatitis virus. Nucleic Acids Res. **12**:3829–3841.
15. **Herman, R. C., S. Adler, R. J. Colonno, A. K. Banerjee, R. A. Lazzarini, and H. Westphal.** 1978. Intervening polyadenylate sequences in RNA transcripts of vesicular stomatitis virus. Cell **15**:587–596.
16. **Herman, R. C., M. Schubert, J. D. Keene, and R. A. Lazzarini.** 1980. Polycistronic vesicular stomatitis virus RNA transcripts. Proc. Natl. Acad. Sci. U.S.A. **77**:4662–4665.
17. **Hopp, T. P., and K. R. Woods.** 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc. Natl. Acad. Sci. U.S.A. **78**:3824–3828.
18. **Huang, Y. T., P. L. Collins, and G. W. Wertz.** 1984. Identification of a new envelope-associated protein of human respiratory syncytial virus, p. 365–368. *In* D. H. L. Bishop and R. W. Compans (ed.), Nonsegmented negative strand viruses, paramyxoviruses and rhabdoviruses. Academic Press, Inc., New York.
19. **Huang, Y. T., and G. W. Wertz.** 1982. The genome of respiratory syncytial virus is a negative-stranded RNA that codes for at least seven mRNA species. J. Virol. **43**:150–157.
20. **Huang, Y. T., and G. W. Wertz.** 1983. Respiratory syncytial virus mRNA coding assignments. J. Virol. **46**:667–672.
21. **Kingsbury, D. W., M. A. Bratt, P. W. Choppin, R. P. Hanson, Y. Hosaka, V. ter Meulen, E. Norrby, W. Plowright, R. Rott, and W. H. Wunner.** 1978. Paramyxoviridae. Intervirology **10**:137–152.
22. **Kozak, M.** 1984. Compilation and analysis of sequences upstream from the translational start sites in eukaryotic mRNAs. Nucleic Acids Res. **12**:857–872.
23. **Kozak, M.** 1984. Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin in vivo. Nature (London) **308**:241–246.
24. **Kozak, M.** 1984. Selection of initiation sites by eucaryotic ribosomes: effect of inserting AUG triplets upstream from the coding sequence for preproinsulin. Nucleic Acids Res. **12**: 3873–3893.
25. **Lamb, R. A., and P. W. Choppin.** 1978. Determination by peptide mapping of the unique polypeptides in Sendai virions and infected cells. Virology **84**:469–478.
26. **Lamb, R. A., and C.-J. Lai.** 1980. Sequence of interrupted and uninterrupted mRNAs and cloned DNA coding for the two overlapping nonstructural proteins of influenza virus. Cell **21**:475–485.
27. **Lamb, R. A., C.-J. Lai, and P. W. Choppin.** 1981. Sequences of mRNAs derived from genome RNA segment 7 of influenza virus—colinear and interrupted mRNAs code for overlapping proteins. Proc. Natl. Acad. Sci. U.S.A. **78**:4170–4174.
28. **Lambert, D. M., and M. W. Pons.** 1983. Respiratory syncytial virus glycoproteins. Virology **130**:204–214.
29. **Land, H., M. Gretz, H. Hauser, W. Lindenmaier, and G. Schutz.** 1981. 5'-Terminal sequences of eukaryotic mRNA can be cloned with high efficiency. Nucleic Acids Res. **9**:2251–2266.
30. **Levine, S.** 1977. Polypeptides of respiratory syncytial virus. J. Virol. **21**:427–431.
31. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning, a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
32. **Maxam, A., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65**:499–560.
33. **Peeples, M., and S. Levine.** 1979. Respiratory syncytial virus polypeptides: their location in the virion. Virology **95**:137–145.
34. **Pringle, C. R., P. V. Shirodaria, H. B. Gimenez, and S. Levine.** 1981. Antigen and polypeptide synthesis by temperature-sensitive mutants of respiratory syncytial virus. J. Gen. Virol. **54**:173–183.
35. **Rose, J. K.** 1980. Complete intergenic and flanking sequences from the genome of vesicular stomatitis virus. Cell **19**:415–421.
36. **Satake, M., and S. Venkatesan.** 1984. Nucleotide sequence of the gene encoding respiratory syncytial virus matrix protein. J. Virol. **50**:92–99.
37. **Shioda, T., Y. Hidaka, T. Kanda, H. Shibuta, A. Namoto, and K. Iwasaki.** 1983. Sequence of 3687 nucleotides from the 3' end of Sendai virus genomic RNA and the predicted amino acid sequences of viral NP, P and C proteins. Nucleic Acids Res. **11**:7317–7331.
38. **Ueba, O.** 1980. Purification and polypeptides of respiratory syncytial virus. Microbiol. Immunol. **24**:361–364.
39. **Venkatesan, S., N. Elango, and R. M. Chanock.** 1983. Construction and characterization of cDNA clones for four respiratory syncytial viral genes. Proc. Natl. Acad. Sci. U.S.A. **80**: 1280–1284.
40. **Wilde, A., and T. Morrison.** 1984. Structural and functional characterization of Newcastle disease virus polycistronic RNA species. J. Virol. **51**:71–76.
41. **Wunner, W. H., and C. R. Pringle.** 1976. Respiratory syncytial virus proteins. Virology **73**:228–243.